# Pathways to Data: From Plans to Datasets

Anastasia Bennett
The Information School
University of Washington
Seattle, Washington
ab233@uw.edu

Will Sutherland
Human Centered Design & Engineering
University of Washington
Seattle, Washington
willtsk@uw.edu

Yubing Tian
The Information School
University of Washington
Seattle, Washington
ytian94@uw.edu

Megan Finn
The Information School
University of Washington
Seattle, Washington
megfinn@uw.edu

Amelia Acker
School of Infomtion
University of Texas at Austin
Austin, Texas
aacker@ischool.utexas.edu

***Abstract*—What is the relationship between Data Management Plans (DMPs), DMP guidance documents, and the reality of end-of-project data preservation and access? In this short paper we report on some preliminary findings of a 3-year investigation into the impact of DMPs on federally funded science in the United States. We investigated a small sample of publicly accessible DMPs (N=14) published using DMPTool. We found that while DMPs followed the National Science Foundation's guidelines, the pathways to the resulting research data are often obscure, vague, or not obvious. We define two "data pathways" as the search tactics and strategies deployed in order to find datasets.**

***Keywords—data management policies, data pathways***

## I. Introduction

Data Management Plans (DMPs) are required by federal research funding agencies such as the National Science Foundation (NSF). Submitted as part of a grant application, they explain the intentions of the principal investigator (PI) for project data management. With the NSF's new data sharing requirements implemented in 2011, DMPs became a key artifact for meeting funding institutions' expectations for data sharing in the US [1]. While this hope has been unevenly realized, current approaches for research data sharing move beyond using DMPs as artifacts to project long-term preservation plans to characterize them as "living" documents that should be revised throughout the course of the project [e.g. [2], [3]]. If data librarians take DMPs seriously as an operational document, then it also makes sense to ask about the connection between DMPs and the resulting research data. That is, how can DMPs help us find data?

We investigated themes from N=14 DMPs dated from 2014-2020, across six NSF directorates that were made with and publicly accessible through DMPTool. This sample serves as a preliminary engagement with DMPs as a resource for examining data management practices. We believe that these plans, which are often occluded from public documentation and peer reviewed publications, reveal emerging new trends that signal a shift in the locus of power over data, not simply within the world of scientific data management and access, but in the development and circulation of scientific knowledge. While prior work has characterized the content and focus of DMPs by assessing internal criteria from performance guidelines [4], in this research we begin to develop a research instrument to connect the DMP, as a data management practice, to resulting datasets. We do this by tracing connections, or pathways, from DMPs to the intended data products they describe.

We begin by surveying literature on data management planning, federally mandated DMP policies in the US, and efforts to evaluate and measure the impact of DMPs on data access. Then we present our method and design for collecting DMPs and following pathways from those documents to their proposed data. Finally, we narratively describe two pathways and discuss what they reveal about scientific data lifecycles.

## II. Related Research

DMPs have been in use to aid data collection and preservation since the 1960s. Information scholars have focused on the role of "data work" *in support* of preservation, access, and ultimately the reproducibility of science [5]. But professional data librarians, sometimes called "information maintainers" [6], data curators, or "data cleaners" [7], have since the 1970s built a professional identity around describing, accessioning, and preserving scientific data in support of access and re-use [8]. While there has long been consensus around the value of data work in the digital library community, resources devoted to associated tasks have lagged. A decade ago, digital library researchers Wallis et al. found that "without external pressure, data management–in the form of digital libraries for capture, curation, and access–is not high on the priority list of science and technology researchers" [8, p. 338]. Thus, within the digital library community, there was growing hope around new data policies such as DMP requirements: "As the idea of data management plans matures due to NIH and NSF requirements, we anticipate greater transparency and common definitions that will form important pieces of future data policy best practices" [9, p. 410].

Anonymized for review.

Since 2011 a number of scholarly communication and digital library researchers have examined the development of institutions' data management requirements, focusing in particular on DMPs' coverage of different aspects of data management as well as different stages of data processing and sharing [9-11]. In a previous study related to this project, our team (ANONYMIZED) examined DMP guideline documents from 15 NSF programs and directorates from the establishment of the policy in 2011 to 2020. Other policy and information science researchers have examined the impact of the federal planning policy on scientific knowledge, data sharing, and institutional repositories [1], [12], [13]. Collectively, these studies suggest that there is a developing, robust literature that examines not just the content of DMPs, but also research that connects DMPs to practices useful for evaluating the plans as well.

## III. Methods

The aim of this ongoing research is to investigate how NSF DMP guidelines have affected the lifecycles of scientific data, knowledge production, and the data management practices of researchers. DMPs are occluded genres of academic documents typically located in the middle of grant applications. They are considered to be personally identifiable information of the research scientists and may contain information that researchers want to keep private at the beginning of their projects. As such, DMPs are hard to find as publicly available resources. Since 2011, the California Digital Library and its founding partners have made DMPTool, a NSF funded project with the goal of creating machine readable DMPs, freely available and allowing researchers to build DMPs using templates and post them publicly (https://dmptool.org/). The corpus of public DMPs contains plans at various stages of completion, across funding agencies, following national and international guidelines.

We employed a sequential two-pronged data collection strategy to locate DMPs in DMPTool for NSF-funded projects and then trace pathways from the projects to their data. First, using various search strategies, a team of six graduate student researchers found DMPs for potentially funded projects in the public corpus (https://dmptool.org/public_plans). The NSF Awards Search (https://www.nsf.gov/awardsearch/) was used to confirm, through triangulation–matching researchers and PIs, similarity of project names, abstracts, and date ranges–a collection of feasible, public DMPs with affiliated NSF funded projects that resulted in non-sensitive data. Our corpus contained 14 DMPs, dated between 2014-2020, from 6 NSF directorates, with the majority coming from Geosciences (8), as well as Biological Sciences (2), Engineering (2), Computer and Information Science and Engineering (1), and Education and Human Resources (1). Aside from the small sample, one major limitation of this study is that we cannot confirm that the public DMPs we collected are the exact documents that were submitted to the NSF as part of final project proposals. DMPTool allows the document creators to publish final DMPs that may be downloaded and edited. However, given the purpose of DMPTool's templates and public hosting, we believe that these DMPs represent close approximations of submitted proposal documents.

We analyzed the DMPs through thematic analysis, capturing relevant details about the pathways to finding project data using online search tactics. From information found in both the DMPs and the NSF awards database, and using various search strategies we tracked down the data affiliated with these projects. Confirmation of datasets from funded projects was again completed by triangulation through online searches matching similarity of PIs, project names, abstracts, described repositories, date ranges, and grant numbers. Additionally, each DMP was evaluated using Section 5 of the DART DMP Rubric Scoresheet (https://osf.io/kh2y6/), which specifically evaluates "plans for data archiving and preservation of access" under a series of six performance criteria. The DART Rubric is a standard for evaluating the content of a DMP and provides a basis of comparison between the plans articulated in the DMP and the situation in which the data was found.

## IV. Findings

Tracing data pathways from DMPs to sites of access and preservation revealed that the DMP was oftentimes insufficient on its own for locating research data. Using the DMP as a starting point, we executed various searches using terms such as the projects' or PIs' names, and repositories stated in the DMP. Yet this method did not guarantee finding the research data's location and it was necessary to supplement information stated in the DMP. We also referred to the projects' NSF award websites to locate publications, as well as PIs' personal websites to obtain additional materials for a given project or publication. Below we present a table that summarizes our findings and two narrative vignettes describing how we used two DMPs to locate research data.

### A. *Section 5 DART Assessment Adaptation*

Table 1 provides an overview of the types of data produced from projects and where they were located if published at the time of this research. We do not provide project names or NSF award numbers here to preserve the confidentiality of the projects' PIs. Our goal is to understand the connections between DMPs and data lifecycles, not to evaluate the compliance of PIs with their proposed plans. The table is an adaptation using Section 5 from the DART DMP Rubric Scoresheet. The scoresheet measures criteria as "Complete/detailed", "Addressed issue, but incomplete", and "Did not address" which we represented as a filled black circle, a half-filled black circle, and an empty circle, respectively. In addition to the six criteria from Section 5, we included the DMP's date, anticipated data types resulting from the project, planned destination of the data, and if we were successful in our data pathways.

### B. *Data Pathway Vignettes*

*a) A Multitude of Data Pathways:* Given the limited treatment of data preservation and archival plans in the DMP–which stated that the data would be available through email listserv, PI's social media, and publication–we were surprised at the multitude of locations this project used to preserve and archive their data. Using the DMP as an entry point together with the project's NSF award website, we found two publications referenced and located both as embedded PDFs on the PI's personal website. In the publications we were able to

TABLE I. SECTION 5 DART ASSESSMENT ADAPTATION

| Plan # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Date of Plan** | 2014 | 2014 | 2014 | 2015 | 2015 | 2015 | 2016 | 2017 | 2017 | 2017 | 2018 | 2018 | 2018 | 2020 |
| **Types of Data from DMPs** | N/A | CSV, PDF | N/A | CSV, KML, TXT | PY, TXT | Dissert-tation | Disser-tation | .R Code, PDF, TXT, Raw Genetic Sequence Reads | PDF, TSV, TXT, XML | PDF, TSV, TXT, XML | PDF, TSV, TXT, XML | N/A | ARCGIS, CSV, Excel, PDF, XLSX, XML, Zip files | N/A |
| **Location of Data (Institution Name)** | UCSD | Arctic Data Center, BCO-DMO | N/A | Hydroshare | Github | Utah State University Digital Commons | Social and Cultural Anthrop-ology Commons | Dropbox, Github, PI website, NCBI Sequence Read Archive | BCO-DMO Woods Hole Open Access Server | BCO-DMO, Woods Hole Open Access Server | BCO-DMO, Woods Hole Open Access Server | Utah State University Digital Commons | Mfield Research and Data Hub, DataOne | DSpace at MIT, JHU IDIES |
| **Successful Data Pathway** | No | Yes | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | No | Yes | No |
| **Provides details about how the data will be archived** | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ◑ | ● | ● | ● |
| **Describes how access to the archived data will be maintained** | ● | ● | ● | ● | ● | ○ | ● | ● | ● | ● | ◑ | ◑ | ● | ◑ |
| **Describes plans for archiving and preserving digital data** | ● | ● | ● | ● | ● | ◑ | ● | ● | ◑ | ◑ | ◑ | ◑ | ● | ◑ |
| **Describes plans for archiving and preserving physical data** | ◑ | ○ | ● | ○ | ○ | ○ | ○ | ◑ | ● | N/A | ○ | ◑ | ● | ◑ |
| **Identifies a timeframe for how long data will be archived** | ● | ● | ● | ● | ◑ | ○ | ● | ◑ | ◑ | ○ | ○ | ○ | ● | ○ |
| **Plan discusses the types or formats of data the investigator expects to retain in their possession** | ● | ● | ● | ● | ◑ | ○ | ◑ | ● | ● | ● | ◑ | ○ | ◑ | ○ |

locate descriptions of where data and supplementary materials were deposited under a section titled “Data Accessibility”. One “raw results” dataset was available through a Dropbox link, presumably maintained by the PI. Meanwhile, another “raw reads” dataset was deposited at the National Center for Biological Information’s Sequence Read Database, that we located using the project ID provided in a publication. According to one of the publications another dataset was deposited at the NOAA’s National Center for Environmental Information with an accession number, but we were unable to successfully locate it despite various searches combining accession number, project name, and PI name. Both publications used GitHub to deposit and preserve various project outputs including: model code, input files, model results, R scripts to recreate analyses and figures, processed datasets, walkthroughs, and even metadata.

*b) The Beeline:* The product of this NSF proposal was expected to be minimal, a couple files of Python source code and a few scripts, less than 1MB total. The DMP stated simply that these products would be stored in a Github repository and all other concerns of data management–access, storage, and archiving–would be managed by the Github platform. There were no publications referenced in the NSF award, so we looked for the investigators' Github repository by typing the name of the proposed tool into a Google search along with the word "Github". The first result returned a Github repository of the same name. The “About” section of this repository stated that it was a NSF funded project aimed at creating the proposed tool. Another Google search result linked to a different archived Github repository by the same individual with a similar name. This archived repository explicitly referenced the NSF award number in its "About" section. Despite the brevity of the DMP and no direct links to the data, it was accessible because of the identifying project name. Confirming it as the correct related data was possible due to the NSF award number referenced in the repository.

## V. DISCUSSION

One of the stated aims of NSF DMP guidance documents is to ensure the discoverability and re-usability of research data. To ensure that this aim is achieved, we find that a well-crafted DMP is oftentimes insufficient. As Table 1 shows, DMPs included specific details regarding the anticipated data types for a given project, suggesting that PIs have a strong sense of what data they will be working with. Though there is less information provided in the DMP with regards to data archiving and preservation, we often found that research data had been deposited, whether that be in institutional repositories or through other methods. For instance, one project (Plan 8) made a dataset accessible through Dropbox, though we are concerned about how stable this is for long-term preservation.

The pathways to data that we reveal were likely not how researchers imagine their data to be found. Given that data

infrastructures are hardly permanent, researchers' plans for *exactly* where their data will live long-term are not always correct, though they do seem to preserve their data regardless. This finding echoes prior literature from our review that found that PIs may be unaware of the labor and skills necessary for long term preservation and access [8]. In some cases, this could be because of poor planning by the PIs, but in other cases repositories themselves change or new repositories become established during the lifetime of the project.

Additionally, we found that whilst individual datasets can be found archived and preserved at institutional data repositories, which are tailored for preserving research data, supplementary materials or "satellite objects"–instrumental to making sense of research data, replicating findings in publications, or providing context to the production of a given dataset–are being stored on Github. We found code, metadata, software, tutorials, and models for replicating research findings and figures deposited there. Github, which was acquired by Microsoft in June 2018, provides a platform for version control via Git and is primarily used for distributed software development [14]. Given the turn to data-driven or data-intensive science and the use of software and models to analyze research data, it is understandable that Github is used by researchers to deposit their code, software packages, and other research materials. However, given its primary function as a software development service, subject to the whims of the platform owners, it may not meet the needs of long-term preservation and archival requirements for research data. Furthermore, as the only linkage we found between data deposited in institutional repositories and Github was through publications or PIs' careful documentation on their personal websites, there is a rupture in the pathway between datasets and the materials required to work with and understand them. Simply put, there are no official or systematic ways to connect datasets in institutional repositories to materials deposited on services such as Github or Dropbox.

## VI. Conclusion

In this study we examine, through DMPs, the planning that is done for research data at the inception of research projects and the realities of end-of-project data archiving, preservation, and providing ongoing accessibility. We found that tracing pathways from DMPs to the locations where the datasets were deposited was difficult as information in the DMPs, as well as project identifiers such as PIs' names or award numbers, were often insufficient. In general, there were no formal and reliable links between awards, PIs, multiple datasets, publications, and software or supplementary materials. These findings reflect the DMP's character as an occluded, static planning document, and highlight some obstacles toward using the DMP as an aid to data discoverability and re-usability.

Future work for this ongoing research project will collect DMPs to add to our corpus and continue analysis of data pathways to better understand the scientific data afterlives across different research areas. We also aim to develop heuristics for analyzing scientific data lifecycles. In addition, future research will revisit the values that DMP guidance policies provide for DMPs and data preservation (ANONYMIZED) and examine whether existing data pathways are commensurate with those values. This investigation of DMPs serves as the first step in understanding emerging data management practices in federally funded science. We endeavor to attend to differences in data management practices across domains and between small and large projects within different research areas.


## Acknowledgment

Anonymized for review.


## References


[1] J. E. Pasek, "Historical Development and Key Issues of Data Management Plan Requirements for National Science Foundation Grants: A Review," *Issues in Science and Technology Librarianship*, 2017, doi: 10.5062/F4QC01RP.

[2] T. Miksa, S. Simms, D. Mietchen, and S. Jones, "Ten principles for machine-actionable data management plans," *PLOS Computational Biology*, vol. 15, no. 3, p. e1006750, Mar. 2019, doi: 10.1371/journal.pcbi.1006750.

[3] J. Chodacki *et al.*, "Implementing Effective Data Practices: Stakeholder Recommendations for Collaborative Research Support," Association of Research Libraries, Sep. 2020. doi: 10.29242/report.effectivedatapractices2020.

[4] A. Whitmire, J. Carlson, B. Westra, P. Hswe, and S. Parham, "The DART Project: using data management plans as a research tool," Oct. 2015, doi: 10.17605/OSF.IO/QH6AD.

[5] G. Downey, K. R. Eschenfelder, and K. Shankar, "Talking About Metadata Labor: Social Science Data Archives, Professional Data Librarians, and the Founding of IASSIST," in *Historical Studies in Computing, Information, and Society: Insights from the Flatiron Lectures*, W. Aspray, Ed. Cham: Springer International Publishing, 2019, pp. 83-113.

[6] A. L. Russell and L. Vinsel, "After Innovation, Turn to Maintenance," *Technology and Culture*, vol. 59, no. 1, 2018, doi: doi:10.1353/tech.2018.0004.

[7] J.-C. Plantin, "Data Cleaners for Pristine Datasets: Visibility and Invisibility of Data Processors in Social Science," *Science, Technology, & Human Values*, vol. 44, no. 1, pp. 52-73, Jan. 2019, doi: 10.1177/0162243918781268.

[8] J. C. Wallis, E. Rolando, and C. L. Borgman, "If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology," *PLOS ONE*, vol. 8, no. 7, p. e67332, Jul. 2013, doi: 10.1371/journal.pone.0067332.

[9] K. A. Bohémier, T. Atwood, A. Kuehn, and J. Qin, "A content analysis of institutional data policies," in *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries - JCDL '11*, Ottawa, Ontario, Canada, 2011, p. 409, doi: 10.1145/1998076.1998159.

[10] D. Dietrich, T. Adamus, A. Miner, and G. Steinhart, "De-Mystifying the Data Management Requirements of Research Funders," 2012, doi: 10.5062/F44M92G2.

[11] M. Williams, J. Bagwell, and M. Nahm Zozus, "Data management plans: the missing perspective," *Journal of Biomedical Informatics*, vol. 71, pp. 130-142, Jul. 2017, doi: 10.1016/j.jbi.2017.05.004.

[12] C. Hudson-Vitale and H. Moulaison Sandy, "Data Management: Plans A Review," *DESIDOC Jl. Lib. Info. Technol.*, vol. 39, no. 06, pp. 322-328, Dec. 2019, doi: 10.14429/djlit.39.06.15086.

[13] J. Thoegersen, "Examination of Federal Data Management Plan Guidelines," *JESLIB*, vol. 4, no. 1, 2015, doi: 10.7191/jeslib.2015.1072.

[14] GitHub, "About archiving content and data on GitHub," 2021. https://docs.github.com/en/github/creating-cloning-and-archiving-repositories/about-archiving-content-and-data-on-github (accessed Mar. 19, 2021).